

Learning Interactions and Relationships between Movie Characters

Anna Kukleva, Makarand Tapaswi, Ivan Laptev

► To cite this version:

Anna Kukleva, Makarand Tapaswi, Ivan Laptev. Learning Interactions and Relationships between Movie Characters. CVPR 2020- IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun 2020, Virtual, United States. 10.1109/CVPR42600.2020.00987 . hal-03017606

HAL Id: hal-03017606

<https://hal.inria.fr/hal-03017606>

Submitted on 21 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Learning Interactions and Relationships between Movie Characters

Anna Kukleva^{1,2}

Makarand Tapaswi¹

Ivan Laptev¹

¹Inria Paris, France

²Max-Planck-Institute for Informatics, Saarbrücken, Germany

akukleva@mpi-inf.mpg.de, {makarand.tapaswi, ivan.laptev}@inria.fr

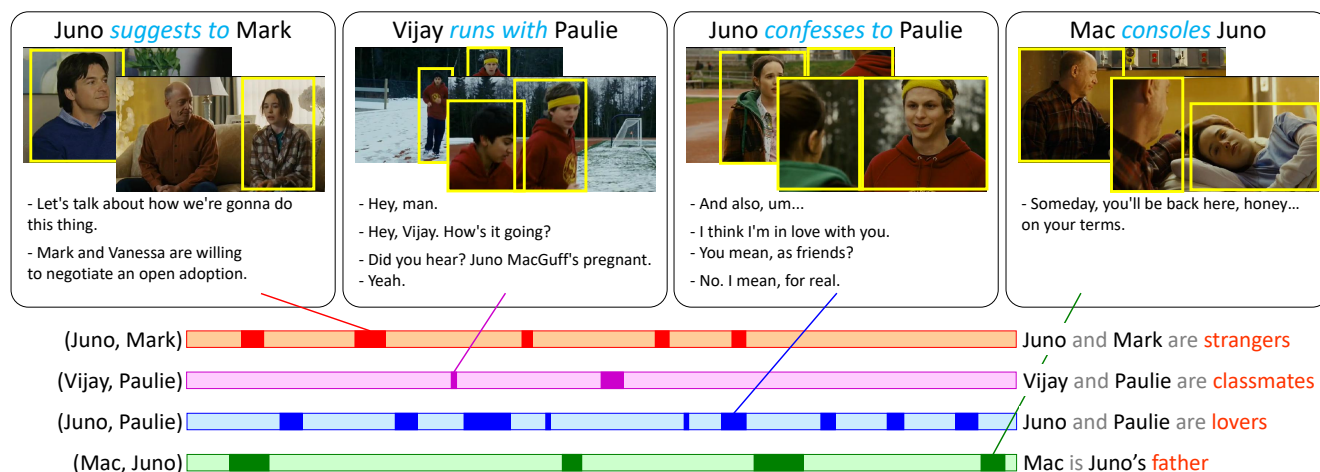


Figure 1: The goal of this work is to jointly predict interactions and relationships between all characters in movies. Some interactions are based on dialog (e.g. suggests, confesses), some are primarily visual (e.g. runs with), and others are based on a fusion of both modalities (e.g. consoles). The colored rows at the bottom highlight when a pair of characters appear in the movie timeline. Their (directed) relationships are presented at the right. Example clips obtained from the movie *Juno*, 2007.

Abstract

Interactions between people are often governed by their relationships. On the flip side, social relationships are built upon several interactions. Two strangers are more likely to greet and introduce themselves while becoming friends over time. We are fascinated by this interplay between interactions and relationships, and believe that it is an important aspect of understanding social situations. In this work, we propose neural models to learn and jointly predict interactions, relationships, and the pair of characters that are involved. We note that interactions are informed by a mixture of visual and dialog cues, and present a multimodal architecture to extract meaningful information from them. Localizing the pair of interacting characters in video is a time-consuming process, instead, we train our model to learn from clip-level weak labels. We evaluate our models on the MovieGraphs dataset and show the impact of modalities, use of longer temporal context for predicting relationships, and achieve encouraging performance using weak labels as compared with ground-truth labels. Code is online.¹

¹<https://annusha.github.io/LIReC/>

1. Introduction

A salient aspect of being human is our ability to interact with other people and develop various relationships over the period of our lives. While some relationships drive the typical interactions experienced by a pair of people in a top-down manner (e.g. parents customarily love and nurture their children); almost all social (non-family) relationships are driven through bottom-up interactions (e.g. strangers become friends over a good chat or a shared drink) [21]. For an intelligent agent to truly be a part of our lives, we will need it to assimilate this complex interplay and learn to behave appropriately in different social situations.

We hypothesize that a first step in this direction involves learning how people interact and what their relationships might be. However, training machines with live, real world, experience-based data is an extremely complicated proposition. Instead, we rely on movies that provide snapshots into key moments of our lives, portraying human behavior at its best and worst in various social situations [47].

Interactions and relationships have been addressed separately in literature. Interactions are often modeled as simple

actions [19, 36], and relationships are primarily studied in still images [29, 43] and recently in videos [30]. However, we believe that a complete understanding of social situations can only be achieved by modeling them jointly. For example, consider the evolution of interactions and relationships between a pair of individuals in a romantic movie. We see that the characters first meet and talk with each other and gradually fall in love, changing their relationship from strangers to friends to lovers. This often leads to them getting married, followed subsequently by arguments or infidelity (a strong bias in movies) and a falling out, which is then reconciled by one of their friends.

The goal of our work is to attempt an understanding of these rich moments of peoples’ lives. Given short clips from a movie, we wish to predict the interactions and relationships, and localize the characters that experience them throughout the movie. Note that our goals necessitate the combination of visual as well as language cues; some interactions are best expressed visually (*e.g.* runs with), while others are driven through dialog (*e.g.* confesses) – see Fig. 1. As our objectives are quite challenging, we make one simplifying assumption - we use *trimmed* (temporally localized) clips in which the interactions are known to occur. We are interested in studying two important questions: (i) can learning to jointly predict relationships and interactions help improve the performance of both? and (ii) can we use interaction and relationship labels at the clip or movie level and learn to identify/localize the pair of characters involved? We refer to this as *weak track prediction*. A solution for the first question is attempted using a multi-task formulation operating on several clips spanning the common pair of characters, while the second uses a combination of max-margin losses with multiple instance learning (see Sec. 3).

Contributions. We conduct our study on 51 movies from the recently released MovieGraphs [47] dataset (see Sec. 4.2). The dataset annotations are based on free-text labels and have long tails for over 300 interaction classes and about 100 relationships. To the best of our knowledge, ours is the first work that attempts to predict interactions and long-term relationships between characters in movies based on visual and language cues. We also show that we can learn to localize characters in the video clips while predicting interactions and relationships using weak clip/movie level labels without a significant reduction in performance.

2. Related Work

We present related work in understanding actions/interactions in videos, studying social relationships, and analyzing movies or TV shows for other related tasks.

Actions and interactions in videos. Understanding actions performed by people can be approached in many different ways. Among them, action classification involves predicting

the dominant activity in a short trimmed video clip [24, 41], while action localization involves predicting the activity as well as temporal extent [15, 39, 51]. An emerging area involves discovering actions in an unsupervised manner by clustering temporal segments across all videos corresponding to the same action class [2, 25, 38].

Recently, there has been an interest in creating large-scale datasets (millions of clips, several hundred classes) for learning actions [1, 5, 11, 18, 34] but none of them reflect person-to-person (p2p) multimodal interactions where several complex actions may occur simultaneously. The AVA challenge and dataset [19] is composed of 15 minute video clips from old movies with *atomic* actions such as pose, person-object interactions, and person-person interactions (*e.g.* talk to, hand wave). However, all labels are based on a short (3 second) temporal window, p2p actions are not annotated between multiple people, and relationship labels are not available. Perhaps closest to our work on studying interactions, Alonso *et al.* [36] predict interactions between two people using person-centered descriptors with tracks. However, the TV-Human Interactions dataset [36] is limited to 4 visual classes in contrast to 101 multimodal categories in our work. As we are interested in studying intricate multimodal p2p interactions and long-range relationships, we demonstrate our methods on the MovieGraphs dataset [47].

Recognizing actions in videos requires aggregation of spatio-temporal information. Early approaches include hand-crafted features such as interest points [26] and Improved Dense Trajectories [48]. With end-to-end deep learning, spatio-temporal 3D Convolutional Neural Networks (*e.g.* I3D [5]) are used to learn video representations resulting in state-of-the-art results on video understanding tasks. For modeling long-videos, learning aggregation functions [16, 33], subsampling frames [49], or accumulating information from a feature bank [50] are popular options.

Relationships in still images. Most studies on predicting social relationships are based on images [14, 17, 29, 40, 43]. For example, the People in Photo Albums (PIPA) [54] and the People in Social Context (PISC) datasets [29] are popular among social relationship recognition. The latter contains 5 relationship types (3 personal, 2 professional), and [29] employs an attention-based model that looks at the entire scene as well as person detections to predict relationships. Alternately, a domain based approach is presented by Sun *et al.* [43] that extends the PIPA dataset and groups 16 social relationships into 5 categories based on Burghental’s theory [4]. Semantic attributes are used to build interpretable models for predicting relationships [43].

We believe that modeling relationships requires looking at long-term temporal interactions between pairs of people, something that still image works do not allow. Thus, our work is fundamentally different from above literature.

Social understanding in videos. Understanding people in

videos goes beyond studying actions. Related topics include clustering faces in videos [22, 45], naming tracks based on multimodal information [35, 44], studying where people look while interacting [13, 32], predicting character emotions [10, 47], modeling spatial relations between objects and characters [27, 42, 55], recognizing actions performed in groups [3, 8], predicting effects for characters [56], producing captions for what people are doing [37, 7], answering questions about events, activities, and character motivations [28, 46, 23], reasoning about social scenes and events [52, 53], understanding social relationships [30, 47], and meta-data prediction using multiple modalities [6].

Perhaps most related to our work on predicting relationships are [30, 31]. Lv *et al.* [31] present the first dataset for modeling relationships in video clips, and propose a multi-stream model to classify 16 relationships. More recently, Liu *et al.* [30] propose a graph network to capture long-term and short-term temporal cues in the video. Different from above works, we address predicting relationships between pairs of characters in an entire movie. We propose a joint model for interactions and relationships as they may influence each other, and also localize the characters in the video.

3. Model

In this section, we present our approach towards predicting the interactions and relationships between pairs of characters (Sec. 3.1), and localizing characters in the video as tracks (Sec. 3.2).

Notation. We define $\mathcal{A}$ as the set of all interaction labels, both visual and spoken (*e.g.* runs with, consoles); and $\mathcal{R}$ as the set of all relationship labels between people (*e.g.* parent-child, friends). We process complete movies, where each movie $\mathcal{M}$ consists of three sets of information:

1. Characters $\mathcal{C}_M = \{c_1, \dots, c_P\}$, each c_j representing a cluster of all face/person tracks for that character.
2. *Trimmed* video clips annotated with interactions $\mathcal{T}_M = \{(v_1, a_1^*, c_{1j}, c_{1k}), \dots, (v_N, a_N^*, c_{Nj}, c_{Nk})\}$, where v_i corresponds to a multimodal video clip, $a_i^* \in \mathcal{A}$ is a directed interaction label, and c_{ij} is used to denote the tracks for character c_j in the clip v_i .
3. Directed relationships between all pairs of characters $\mathcal{R}_M = \{r_{jk}^i = \text{relationship}(v_i, c_j, c_k)\}$ for all clips $i \in [1, N]$. For simplicity of notation, we assign a relationship label r_{jk}^i to each clip. However, note that relationships typically span more than a clip, and often the entire movie (*e.g.* parent-child).

For each clip v_i , our goal is to predict the primary interaction a_i , the characters c_{ij} and c_{ik} that perform this interaction, and their relationship r_{jk}^i . In practice, we process several clips belonging to the same pair of characters as predicting relationships with a single short clip can be

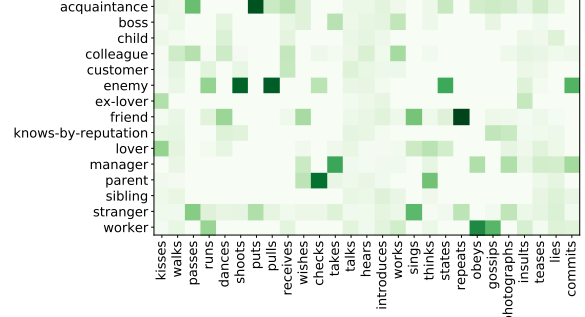


Figure 2: Normalized correlation map between (selected) interactions and relationships. Darker regions indicate higher scores.

quite challenging, and using multiple clips helps improve performance.

We denote the *correct* pair of characters in a tuple $(v_i, a_i^*, c_{ij}, c_{ik})$ from $\mathcal{T}$ as $p_i^* = (c_{ij}, c_{ik})$, and the set of all character pairs as $\mathcal{P}_M = \{(c_j, c_k) \forall j, k, j \neq k\}$.

Note that the interaction tuples in $\mathcal{T}$ may be temporally overlapping with each other. For example, Jack may *look at* Jill while she *talks to* him. We deal with such interaction labels from overlapping clips in our learning approach by masking them out in the loss function.

3.1. Interactions and Relationships in a Clip

Fig. 2 shows example correlations between a few selected interactions and all 15 relationships in our dataset. We observe that interactions such as *obeys* go together with *worker-manager* relationships, while an *enemy* may *shoot*, or *pull* (a weapon), or *commit* a crime. Motivated by these correlations, we wish to learn interactions and relationships jointly.

When the pair of characters that interact is known, we predict their interactions and relationships using a multi-modal clip representation $\Phi(v_i, p_i^*) \in \mathbb{R}^D$. As depicted in Fig. 3, Φ combines features from multiple sources such as visual and dialog cues from the video, and character representations by modeling their spatio-temporal extents (via tracking).

Interactions. We use a two-layer MLP with a classification layer $\mathbf{W}^{I2} \in \mathbb{R}^{|\mathcal{A}| \times D}$, $\mathbf{b}^{I2} \in \mathbb{R}^{|\mathcal{A}|}$ to predict interactions between characters. The score for an interaction a in clip v is computed as

$$s_I(v, a) = \sigma_I(\mathbf{w}_a^{I2} \cdot \text{ReLU}(\mathbf{W}^{I1} \Phi(v, p^*) + \mathbf{b}^{I1}) + b_a^{I2}), \quad (1)$$

where $\sigma(\cdot)$ represents the sigmoid operator. We learn the clip representation parameters along with the MLP by minimizing the max-margin loss function for each sample

$$L_I(v) = \sum_{\substack{\bar{a} \in \mathcal{A} \setminus \mathcal{O}_I(v) \\ \bar{a} \neq a^*}} [m_I - s_I(v, a^*) + s_I(v, \bar{a})]_+, \quad (2)$$

where $[\cdot]_+$ is a ReLU operator, m_I is the margin, and $\mathcal{O}_I(v)$ is the set of interaction labels from clips temporally over-

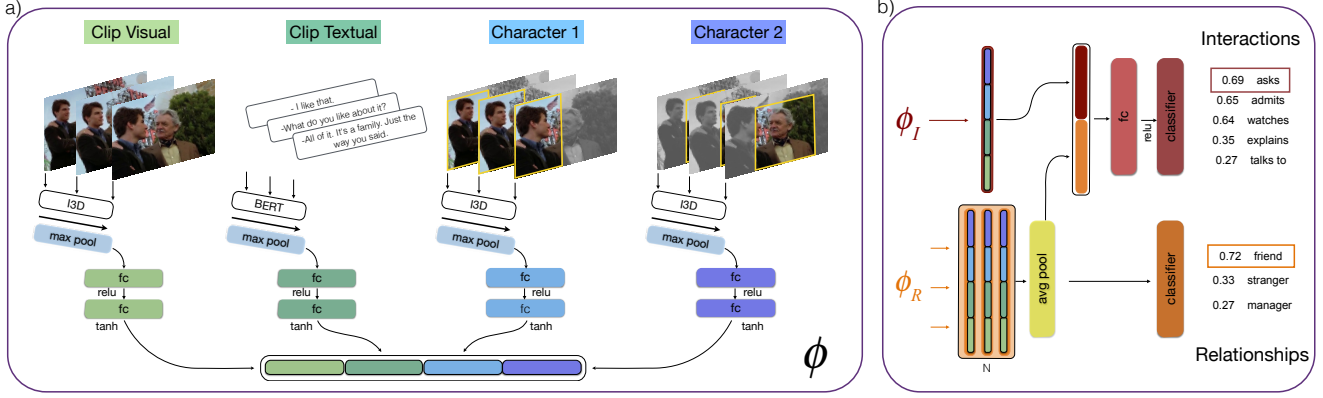


Figure 3: Model architecture. **Left:** Our input is a trimmed video clip for one interaction, and consists of visual frames and all dialogues within its duration. Each interaction is associated with two characters, and they are represented visually by extracting features from cropped bounding boxes. Modalities are processed using fixed pre-trained models (BERT for textual, I3D for visual) to extract clip representations denoted by $\Phi(v)$. **Right:** In the second panel, we show the architecture of our joint interaction and relationship prediction model. In particular, multiple clips are used to compute relationships, and we fuse these features while computing interaction labels.

lapping with v . The loss encourages our model to associate the correct interaction a^* with the clip v , while pushing other non-overlapping interaction labels $\bar{a}$ away. During inference, we predict the interaction for a clip v as $\hat{a} = \arg \max_a s_I(v, a)$.

Relationships. While interactions are often short in duration (few seconds to a minute), relationships in a movie may last for several minutes to the entire movie. To obtain robust predictions for relationships, we train a model that observes several trimmed video clips that portray the same pair of characters. Let us denote $V_{jk} \subset \{v_1, \dots, v_N\}$ as one such subset of clips that focus on characters c_j and c_k . In the following, we drop the subscripts jk for brevity.

Similar to predicting interactions, we represent individual clips of V using Φ , apply a pooling function $g(\cdot)$ (e.g. avg, max) to combine the individual clip representations as $\Phi(V, p^*) = g(\Phi(v, p^*))$ and adopt a linear classifier $\mathbf{W}^R \in \mathbb{R}^{|\mathcal{R}| \times D}$, $\mathbf{b}^R \in \mathbb{R}^{|\mathcal{R}|}$ to predict relationships. The scoring function

$$s_R(V, r) = \sigma_r(\mathbf{w}_r^R \Phi(V, p^*) + b_r^R) \quad (3)$$

computes the likelihood of character pair p^* from the clips V having relationship r . We train model parameters using a similar max-margin loss function

$$L_R(V) = \sum_{\substack{\bar{r} \in \mathcal{R} \\ \bar{r} \neq r^*}} [m_R - s_R(V, r^*) + s_R(V, \bar{r})]_+, \quad (4)$$

that attempts to score the correct relationship r^* higher than others $\bar{r}$. Unlike L_I , we assume that a single label applies to all clips in V . If a pair of characters change relationships (e.g. from *strangers* to *friends*), we select the set of clips V during which a single relationship is present. At test time, we predict the relationship as $\hat{r} = \arg \max_r s_R(r, V)$.

Joint prediction of interactions and relationships is performed using a multi-task formulation. We consider multiple clips V and train our model to predict the relationship as well as all interactions of the individual clips jointly. We introduce a dependency between the two tasks by concatenating the clip representations for interactions $\Phi_I(v, p^*)$ and relationships $\Phi_R(V, p^*)$. Fig. 3 visualizes the architecture used for this task. We predict interactions as follows:

$$s_I(v, V, a) = \sigma(\mathbf{w}_a^{I2} \cdot \text{ReLU}(\mathbf{W}^{I1}[\Phi_I(v, p^*); \Phi_R(V, p^*)])). \quad (5)$$

Linear layers include biases, but are omitted for brevity. The loss function $L_I(v)$ now uses $s_I(v, V, a)$, but remains unchanged otherwise. The combined loss function is

$$L_{IR}(V) = L_R(V) + \frac{\lambda}{|V|} \sum_{v \in V} L_I(v), \quad (6)$$

where λ balances the two losses.

3.2. Who is interacting?

Up until now, we assumed that a clip v portrays two *known* characters that performed interaction a . However, movies (and the real world) are often more complex, and we observe that several characters may be interacting simultaneously. To obtain a better understanding of videos, we present an approach to predict the characters along with the interactions they perform, and their relationship.

While the interaction or relationship may be readily available as a clip-level label, localizing the pair of characters in the video can be a tedious task as it requires annotating tracks in the video. We present an approach that can work with such *weak* (clip-level) labels, and estimate the pair of characters that may be interacting.

Predicting interactions and characters. As a first step, we look at jointly predicting interactions and the pair of charac-

ters. Recall that p_i^* denotes the *correct* pair of characters in a clip tuple consisting of v_i , and $\mathcal{P}_M$ is the set of all character pairs in the movie. We update the scoring function (Eq. 1) to depend on the chosen pair of characters $p \in \mathcal{P}_M$ as

$$s_{IC}(v, a, p) = \sigma(\mathbf{w}_a^{I2} \cdot \text{ReLU}(\mathbf{W}^{I1} \Phi(v, p))), \quad (7)$$

where $\Phi(v, p)$ now encodes the clip representation for any character pair p (we use zeros for unseen characters). We train our model to predict interactions and the character pair by minimizing the following loss

$$L_{IC}(v) = \sum_{\substack{\bar{a} \in \mathcal{A} \setminus \mathcal{O}_I(v) \\ \bar{p} \in \mathcal{P}_M \\ (\bar{a}, \bar{p}) \neq (a^*, p^*)}} [m_{IC} - s_{IC}(v, a^*, p^*) + s_{IC}(v, \bar{a}, \bar{p})]_+ \quad (8)$$

If we consider the scoring function $s_{IC}(v, a, p)$ as a matrix of dimensions $|\mathcal{P}_M| \times |\mathcal{A}|$, the negative samples are taken from everywhere except columns that have an overlapping interaction label $\mathcal{O}_I(v)$, and the element where $(\bar{a} = a^*, \bar{p} = p^*)$. At test time, we compute the character pair prediction accuracy given ground-truth (GT) interaction, interaction accuracy given GT character pair, and joint accuracy by picking the maximum score along both dimensions.

Training with weak labels. When the GT character pair p^* is *not known* during training, we modify the loss from Eq. 8 by first choosing the pair $\hat{p}^*$ that scores highest for the current parameters and a^* , that is known during training.

$$\hat{p}^* = \arg \max_p s_{IC}(v, a^*, p), \quad (9)$$

$$L_{IC}^{\text{weak}}(v) = \sum_{\substack{\bar{a} \in \mathcal{A} \setminus \mathcal{O}_I(v), \bar{a} \neq a^* \\ \bar{p} \in \mathcal{P}_M}} [m_{IC} - s_{IC}(v, a^*, \hat{p}^*) + s_{IC}(v, \bar{a}, \bar{p})]_+ \quad (10)$$

In contrast to the case when we know GT p^* , we discard negatives from the entire column ($a = a^*$) to prevent minor changes in choosing $\hat{p}^*$ from suppressing other character pairs. In practice, we treat $s_{IC}(v, a^*, p)$ as a multinomial distribution and sample $\hat{p}^*$ from it to prevent the model from getting stuck at only one pair. Inference is performed in a similar way as above.

Hard negatives. Training a model with max-margin loss can affect performance if the loss is satisfied ($= 0$) for most negative samples. As demonstrated in [12], choosing hard negatives by using max instead of $\sum$ can help improve performance. We adopt a similar strategy for selecting hard negatives, and compute the loss over all character pairs with their best interaction, *i.e.* $\sum_{\bar{p} \in \mathcal{P}_M} \max_{\bar{a}}(\cdot)$ in Eq. 8 and 10.

Prediction of interactions, relationships, and characters. We present the loss function used to learn a model that jointly estimates which characters are performing what interactions and what are their relationships. Similar to Eq. 7, we first modify the relationship score to depend on p :

$$s_{RC}(V, r, p) = \sigma(\mathbf{w}_r^R g(\Phi(V, p)) + b_r^R). \quad (11)$$

This is used in a weak label loss function similar to Eq. 10.

$$\hat{p}^* = \arg \max_p s_{RC}(V, r^*, p) + s_{IC}(v, a^*, p), \quad (12)$$

$$L_{RC}^{\text{weak}}(V) = \sum_{\substack{\bar{r} \in \mathcal{R}, \bar{r} \neq r^* \\ \bar{p} \in \mathcal{P}_M}} [m_{RC} - s_{RC}(V, r^*, \hat{p}^*) + s_{RC}(V, \bar{r}, \bar{p})]_+, \quad (13)$$

$$L_{IRC}^{\text{weak}}(V) = L_{RC}^{\text{weak}}(V) + \frac{\lambda}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} L_{IC}^{\text{weak}}(v). \quad (14)$$

During inference, we combine the scoring functions s_{IC} and s_{RC} to produce a 3D tensor in $|\mathcal{P}_M| \times |\mathcal{A}| \times |\mathcal{R}|$. As before, we compute character pair accuracy given GT a^* and r^* , interaction accuracy given GT p^* and r^* , and relationship accuracy given GT p^* and a^* . We are also able to make joint predictions on all three by picking the element that maximizes the tensor over all three dimensions.

4. Experiments

We start by describing implementation details (Sec. 4.1), followed by a brief analysis of the dataset and the challenging nature of the task (Sec. 4.2). In Sec. 4.3 we examine interaction and relationship prediction performance as separate and joint tasks. Sec. 4.4 starts with learning interactions and estimating the pair of characters simultaneously. Finally, we also discuss predicting relationships jointly with interactions and localizing character pairs. We present both quantitative and qualitative evaluation throughout this section.

4.1. Implementation Details

Visual features. We extract visual features for all clips using a ResNeXt-101 model [20] pre-trained on the Kinetics-400 dataset. A batch of 16 consecutive frames is encoded, and feature maps are global average-pooled for the clip representation, and average pooled over a region of interest (ROI Pool) to represent characters. Given a trimmed clip v_i , we max pool above extracted features over the temporal span of the clip to pick the most important segments.

Dialog features. To obtain a text representation, all dialogues are first parsed into sentences. A complete sentence may be as short as a single word (*e.g.* “Hi.”) or consist of several subtitle lines. Multiple lines are also joined if they end with “...”. Then, each sentence is processed using pre-trained BERT-base model with a masked sentence from the next person if it exists. We supply a mask for every second sentence as done in the sentence pair classification task (for more details, *c.f.* [9]). We max pool over all sentences uttered in a trimmed clip to obtain a final representation.

Note that every clip always has a visual representation. In the absence of dialog or tracks, we set the representations for missing modalities to 0.

Clip representation Φ . We process the feature vector corresponding to each modality obtained after max pooling over

the temporal extent with a two-layer MLP. Dropout (with $p = 0.3$), ReLU and $\tanh(\cdot)$ non-linearities are used in the MLP. The final clip representation is a concatenation of all modalities (see Fig. 3 left).

Multi-label masking. As multiple interactions may occur at the same time or have overlapping temporal extents with other clips, we use masking to exclude negative contributions to the loss function by such labels. $\mathcal{O}_I(v)$, the labels corresponding to the set of clips overlapping with v , are created by checking for an overlap (IoU) greater than 0.2.

Learning. We train our models with a batch size of 64, and use the Adam optimizer with a learning rate of $3e-5$.

4.2. Dataset

We evaluate our approach on the MovieGraphs dataset [47]. The dataset provides detailed graph-based annotations of social situations for 7600 scenes in 51 movies. Two main types of interactions are present—*detailed interactions* (e.g. laughs at) last for a few seconds and are often a part of an overarching *summary interaction* (e.g. entertains) that may span up to a minute. We ignore this distinction for this work and treat all interactions in a similar manner. These hierarchical annotations are a common source of multiple labels being associated with the same timespan in the video.

The total number of interactions is different from the number of p2p instances as some interactions involve multiple people. For example, in an interaction where a couple (c_j and c_k) *listens to* their therapist (c_l), two p2p instances are created: $c_j \rightarrow \text{listens to} \rightarrow c_l$ and $c_k \rightarrow \text{listens to} \rightarrow c_l$.

The dataset is partitioned into train (35 movies), validation (7 movies) and test (9 movies) splits. The train set consists of 15,516 interactions (and 20,426 p2p instances) and 2,676 pairs of people with annotated relationships. Validation and test sets have 3,992 and 5,380 p2p instances respectively, and about 600 relationship pairs each.

Missing labels. A relationship label is available for 64% of the interactions in which at least two people participate. For a pair of people associated with an interaction, both have track features for 76% of the dataset. In other cases one or no characters appear due to failure in tracking or being out of the scene. For evaluation, we only consider samples that have a relationship, or when a pair of characters appear.

Merging interaction and relationship labels. We reduce the number of interaction labels from 324 to 101, and relationships from 106 to 15 to mitigate severe problems of long tail with only 1-3 samples per class. However, the merging does not adversely affect the diversity of classes, e.g. *reassures*, *wishes*, *informs*, *ignores* are different interactions in our label set related to communication.

We adopt a hierarchical approach to merge interactions. Firstly, all classes are divided into 4 categories: (i) informative or guiding (e.g. *explains*, *proposes*, *assists*, *guide*) (ii)

Modalities			Interaction Accuracy		
Visual	Dialog	Tracks	Top-1	Top-1 Soft	Top-5
✓	-	-	18.7	24.6	45.8
-	✓	-	22.4	30.1	50.6
✓	✓	-	25.0	31.9	54.8
✓	✓	✓	26.1	32.6	57.3

Table 1: Interaction prediction accuracy for different modalities.

involving movement (e.g. *hits*, *plays*, *embraces*, *catches*); (iii) neutral valence (e.g. *avoids*, *pretends*, *reads*, *searches*); and (iv) negative valence (e.g. *scolds*, *mocks*, *steals*, *complaints*). Within each of the subclasses we merge interactions based on how similar their meanings are in common usage – this process is verified by multiple people.

We also reduce the number of relationships to 15 major classes: stranger, friend, colleague, lover, enemy, acquaintance, ex-lover, boss, worker, manager, customer, knows-by-reputation, parent, child and sibling.

Directed interactions and relationships are used between one person to another. For example when a *parent* $\rightarrow$ *informs* $\rightarrow$ *child*, the opposite directed interaction from the child to their parent can be *listens to* or *ignores*. Additionally, interactions and relationships can also be bidirectional, both people act with the same intention e.g. *lovers kiss* each other.

4.3. Predicting Interactions and Relationships

We first present results for predicting interactions and relationships separately, followed by our joint model.

Interaction classification. We analyze the influence of each modality for interaction classification separately in Table 1. Dialogs have a stronger impact on model performance as compared to visual features owing to the prevalence of conversation based interactions in movies. However, both modalities are complementary and when taken together provide a 2.6% increase in accuracy. As expected, combining all modalities including tracks for each participating character provide the highest prediction accuracy at 26.1%.

Apart from accuracy, we report *soft accuracy*, a metric that treats a prediction as correct when it matches any of the interactions overlapping with the clip, i.e. $\hat{a} \in a^* \cup \mathcal{O}_I(v)$. When using all modalities, we achieve 32.6% accuracy.

In Fig. 4 we see two example interactions that are challenging to predict based on visual cues alone. In the top example, we see that the ground-truth label *reads* is emphasized, possibly due to the dialog mentioning letters, and is chosen with highest score upon examining the visual tracks. The bottom example is an interesting case where no dialog (all 0 vector) helps predictions. In this case, our model seems to have learned that *leaving* corresponds to *walking* without any dialog. Again, by including information about tracks, our model is able to predict the correct label.

We also investigate the influence of different temporal feature aggregation methods in Table 4. Max-pooling outper-

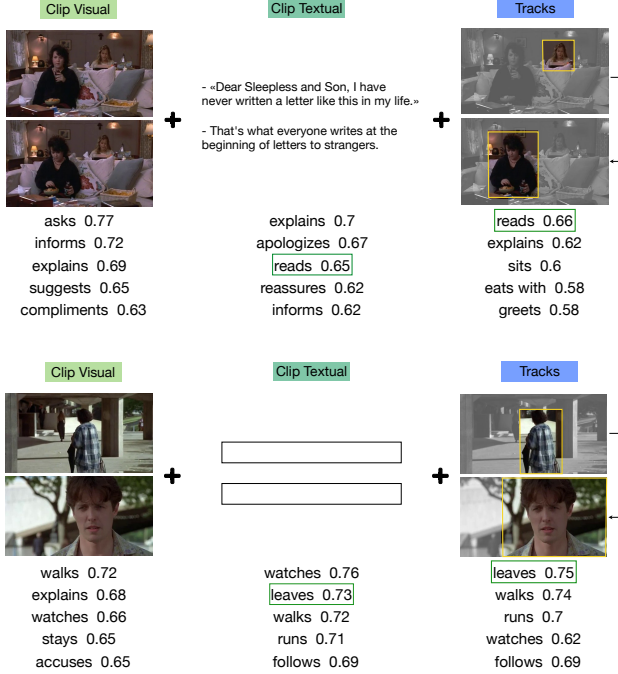


Figure 4: Influence of different modalities on interaction prediction performance. In each example, we show the top 5 interactions predicted by the visual cues (left), the visual + dialog cues (center), and visual + dialog + track information (right). The correct label is marked with a green bounding rectangle. Discussion in Sec. 4.3.

forms both average and sum as it allows to form the clip-level representations including the most influential segments.

Relationship classification. Relationships are often consistent for long durations in a movie. For example, *strangers* do not become *friends* in one moment, and *parents* always stay *parents*. We hypothesize that it is challenging to predict a relationship by watching one interaction, and show the impact of varying the number of clips (size of V) in Fig. 5. We see that the relationship accuracy improves steadily as we increase the number of clips. The drop at 6 clips is within variance. We choose 18 clips as a trade-off between performance and speed. During training, we randomly sample up to 18 clips for the same pair of people having the same relationship. At test time, the clips are fixed and uniformly distributed along all occurrences of pairs of character.

Joint prediction for interactions and relationships. We set the loss trade-off parameter $\lambda = 1.5$ and jointly optimize the network to predict interactions and relationships. We evaluate different options on how the two tasks are modeled

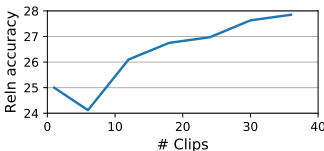


Figure 5: Relationship accuracy increases as we analyze more clips. We choose 18 clips as a trade-off between performance and speed.

Task	Random	Int. only	Rel. only	Joint
Interaction	0.99	26.1	-	26.3
Relationship	6.67	-	26.8	28.1

Table 2: Top-1 accuracy for the joint prediction of Int. and Rel.

Method	Int.	Rel.	Method	Int.
Rel. $\leftrightarrow$ Int.	25.3	26.8	avg	24.2
Rel. $\leftarrow$ Int.	26.3	25.9	sum	25.4
Rel. $\rightarrow$ Int.	26.3	28.1	max	26.1

Table 3: Different architectures for joint modeling of interactions and relationships.

Table 4: Impact of temporal aggregation methods on interaction accuracy.

Supervision	Negatives	Multinom. Sampling	Accuracy Int.	Accuracy Character	Joint
Random	-	-	0.99	15.42	0.15
Full	sum	-	25.5	55.0	14.2
Weak	sum	-	18.9	20.0	4.6
Weak	sum	✓	25.1	25.0	7.8
Weak	sum-max	✓	23.0	32.3	8.2

Table 5: Joint prediction of **interactions** and **character pairs** for fully and weakly supervised settings. See Sec. 4.4 for a discussion.

jointly in Table 3. Overall, concatenating relationship feature for predicting interactions performs best (Rel. $\rightarrow$ Int.). Table 2 shows that the relationship accuracy improves by 1.3%, while interactions gain a meagre 0.2%.

On further study, we observe that some interactions achieve large improvements, while others see a drop in performance. For example, interactions such as *hugs* (+17%), *introduces* (+14%), and *runs* (+12%), are associated with specific relationships: *siblings*, *child*, *lover* with *hugs*; *enemy*, *lover* with *runs*. On the other hand, a few other interactions such as *talks to*, *accuses*, *greet*, *informs*, *yells* see a drop in performance from 1-8%, perhaps as they have the same top-3 relationships: *friend*, *colleague*, *stranger*.

Relationships show a similar trend. *Sibling*, *acquaintance*, *lover* correspond to specific interactions such as *hugs*, *greet*, *kisses* and improve by 11%, 8%, and 7% respectively. While *boss* and *manager* have rather generic interactions *asks*, *orders*, *explains* and reduce by 5-7%.

We observe that joint learning does help improve performance. However, interactions performed by people with common relationships, or relationships that exhibit common interactions are harder for our joint model to identify leading to small overall improvement. We believe this is made harder due to the long tail classes.

4.4. Localizing Characters

We present an evaluation of character localization and interaction prediction in Table 5. We report **interaction**

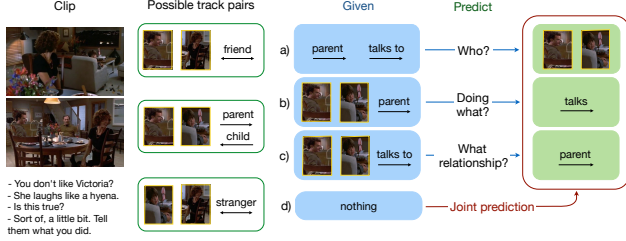


Figure 6: Example for joint prediction of interaction (Int), relationship (Rel), and character pair (Char) by our model. The visual clip, dialog, and possible track pairs are presented on the left. Given 2 pieces of information, we are able to answer the third: *Who?* Int + Rel $\rightarrow$ Char; *Doing what?* Char + Rel $\rightarrow$ Int; and *What relationship?* Char + Int $\rightarrow$ Rel. We can also jointly predict all three components by maximizing scores along all dimensions of the 3D tensor. Best seen on screen with zoom.

Supervision	Negatives	Multinom. Sampling	Accuracy			
			Int.	Rel.	Char.	Joint
Random	-	-	0.99	6.67	15.42	0.01
Full	sum	-	25.8	16.6	88.3	2.71
Weak	sum	✓	25.8	12.0	42.0	0.86
Weak	sum-max	✓	20.8	21.8	33.9	2.14

Table 6: Joint **interaction**, **relationship**, and **character pair** prediction accuracy. Other labels are provided when predicting columns: Int., Rel., and Char. See Sec. 4.4 for a discussion.

accuracy given the correct character pair; **character pair** prediction accuracy given the correct interaction; and the overall accuracy as **joint**.

Training with full supervision. In the case when the pair of characters are known (ground-truth pair p^* is given), we achieve 25.5% accuracy for interactions. This is comparable to the setting where we only predict interactions (at 26.1%). We believe that the difference is due to our goal to maximize the score for the correct interaction and character pair over the entire matrix $|\mathcal{P}_M| \times |\mathcal{A}|$. The joint accuracy is 14.2%, significantly higher than random at 0.15%.

Training with weak supervision. Here, interaction labels are applicable at the clip-level, and we are unaware of which characters participate in the interaction even during training. Table 5 shows that sampling a character pair is better than $\arg \max$ in Eq. 9 (4.6% vs. 7.8% joint accuracy) as it prevents the model from getting stuck at a particular selection. Furthermore, switching training from sum over all negatives to hard negatives (sum-max) after a burn-in period of 20 epochs further improves accuracy to 8.2%.

Joint character localization, interaction and relationship prediction is presented in Table 6. In the case of learning with GT character pairs (fully supervised), including learning of relationships boosts accuracy for predicting character pairs to an impressive 88.3%. The interaction accuracy also

All methods	Supervision	Accuracy			
		Int.	Rel.	Char.	Joint
Int only	-	20.7	-	-	-
Rel only	-	-	22.4	-	-
Int + Rel	-	20.7	20.5	-	-
Int + Char	Full	19.7	-	52.8	11.1
Int + Char	Weak	17.9	-	29.7	6.34
Int + Rel + Char	Full	20.0	18.6	88.8	2.29
Int + Rel + Char	Weak	15.6	29.6	21.6	1.50

Table 7: Evaluation on the test set. The columns Int., Rel, and Char refer to interaction, relationship, and character pair prediction accuracy. During joint learning with full/weak supervision, individual accuracies are reported when other labels are given.

increases to 25.8% as compared against 25.5% when training without relationships (Table 5).

When learning with weak labels, we see similar trends as before. Both multinomial sampling and switching from all (sum) to hard (sum-max) negatives improves the joint accuracy to a respectable 2.14% as compared to 2.71% with full supervision. Fig. 15 shows an example prediction from our dataset. We present joint prediction when no information is provided in part d in contrast to parts a, b, c where two of three pieces of information are given.

Test set. Table 7 compiles results of all our models on the test set. We see similar trends, apart from a drop in relationship accuracy due to different val and test distributions.

Overall, we observe that learning interactions and relationships jointly helps improve performance, especially for classes that have unique correspondences, but needs further work on other categories. Additionally, character localization is achievable and we can train models with weak labels without significant drop in performance.

5. Conclusion

We presented new tasks and models to study the interplay of interactions and relationships between pairs of characters in movies. Our neural architecture efficiently encoded multimodal information in the form of visual clips, dialog, and character pairs that were demonstrated to be complementary for predicting interactions. Joint prediction of interactions and relationships was found to be particularly beneficial for some classes. We also presented an approach to localize character pairs given their interaction/relationship labels at a clip-level, *i.e.* without character-level supervision during training. We will share modifications made to the MovieGraphs dataset to promote future work in this exciting area of improving understanding of human social situations.

Acknowledgements. This project was partially supported by the Louis Vuitton - ENS Chair on Artificial Intelligence and the French government under Agence Nationale de la Recherche as part of “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

- [1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.
- [2] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *CVPR*, 2016.
- [3] Timur Bagautdinov, Alexandre Alahi, Francois Fleuret, Pascal Fua, and Silvio Savarese. Social Scene Understanding: End-To-End Multi-Person Action Localization and Collective Activity Recognition. In *CVPR*, 2017.
- [4] Daphne B. Burgental. Acquisition of the Algorithms of Social Life: A Domain-Based Approach. *Psychological Bulletin*, 126(2):187–219, 2000.
- [5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.
- [6] Paola Cascante-Bonilla, Kalpathy Sitaraman, Mengjia Luo, and Vicente Ordonez. Moviescope: Large-scale analysis of movies using multiple modalities. *ArXiv*, abs/1908.03180, 2019.
- [7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *ECCV*, 2018.
- [8] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *CVPR*, 2016.
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [10] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. Emotion recognition in the wild challenge 2013. In *International Conference on Multimodal Interactions*, 2013.
- [11] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jurgen Gall, Rainer Stiefelhofen, and Luc Van Gool. Holistic large scale video understanding. *arXiv preprint arXiv:1904.11451*, 2019.
- [12] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *BMVC*, 2018.
- [13] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *CVPR*, 2019.
- [14] Andrew C Gallagher and Tsuhan Chen. Understanding images of groups of people. In *CVPR*, 2009.
- [15] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, 2019.
- [16] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. ActionVLAD: Learning spatio-temporal aggregation for action classification. In *CVPR*, 2017.
- [17] Arushi Goel, Keng Teck Ma, and Cheston Tan. An end-to-end network for generating social relationship graphs. In *CVPR*, 2019.
- [18] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The “something something” video database for learning and evaluating visual common sense. In *ICCV*, 2017.
- [19] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018.
- [20] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, 2018.
- [21] R. A. Hinde. Interactions, Relationships and Social Structure. *Man*, 11(1):1–17, 1976.
- [22] SouYoung Jin, Hang Su, Chris Stauffer, and Erik Learned-Miller. End-to-end Face Detection and Cast Grouping in Movies using ErdsRnyi Clustering. In *ICCV*, 2017.
- [23] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. PororoQA: A Cartoon Video Series Dataset for Story Understanding. In *NIPS Workshop on Large Scale Computer Vision Systems*, 2016.
- [24] Hilde Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: A Large Video Human Motion Recognition. In *ICCV*, 2011.
- [25] Anna Kukleva, Hilde Kuehne, Fadime Sener, and Jurgen Gall. Unsupervised learning of action classes with continuous temporal embedding. In *CVPR*, 2019.
- [26] Ivan Laptev and Tony Lindeberg. Space-time Interest Points. In *ICCV*, 2003.
- [27] Benjamin Laxton, Jongwoo Lim, and David Kriegman. Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In *CVPR*, 2007.
- [28] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. TVQA: Localized, Compositional Video Question Answering. In *EMNLP*, 2018.
- [29] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Dual-glance model for deciphering social relationships. In *CVPR*, 2017.
- [30] Xinchun Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan, and Tao Mei. Social relation recognition from videos via multi-scale spatial-temporal reasoning. In *CVPR*, 2019.
- [31] Jinna Lv, Wu Liu, Lili Zhou, Bin Wu, and Huadong Ma. Multi-stream Fusion Model for Social Relation Recognition from Videos. In *International Conference on Multimedia Modeling*, 2018.
- [32] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net: revisiting people looking at each other in videos. In *CVPR*, 2019.

- [33] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv:1706.06905*, 2017.
- [34] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfrueud, Carl Vondrick, and Aude Oliva. Moments in time dataset: one million videos for event understanding. *PAMI*, 2019.
- [35] Arsha Nagrani and Andrew Zisserman. From Benedict Cumberbatch to Sherlock Holmes: Character Identification in TV series without a Script. In *BMVC*, 2017.
- [36] Alonso Patron, Marcin Marszalek, Andrew Zisserman, and Ian Reid. High five: Recognising human interactions in tv shows. In *BMVC*, 2010.
- [37] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *IJCV*, 2017.
- [38] Fadime Sener and Angela Yao. Unsupervised learning and segmentation of complex activities from video. In *CVPR*, 2018.
- [39] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-deconvolutional networks for precise temporal action localization in untrimmed videos. In *CVPR*, 2017.
- [40] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In *NIPS*, 2013.
- [41] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv:1212.0402*, 2012.
- [42] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In *ECCV*, 2018.
- [43] Qianru Sun, Bernt Schiele, and Mario Fritz. A domain based approach to social relation recognition. In *CVPR*, 2017.
- [44] Makarand Tapaswi, Martin Buml, and Rainer Stiefelham. “Knock! Knock! Who is it?” Probabilistic Person Identification in TV series. In *CVPR*, 2012.
- [45] Makarand Tapaswi, Marc T. Law, and Sanja Fidler. Video Face Clustering with Unknown Number of Clusters. In *ICCV*, 2019.
- [46] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelham, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 2016.
- [47] Paul Vicol, Makarand Tapaswi, Lluís Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *CVPR*, 2018.
- [48] Heng Wang and Cordelia Schmid. Action Recognition with Improved Trajectories. In *ICCV*, 2013.
- [49] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.
- [50] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, 2019.
- [51] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. *ICCV*, 2017.
- [52] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence. In *CVPR*, 2019.
- [53] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From Recognition to Cognition: Visual Commonsense Reasoning. In *CVPR*, 2019.
- [54] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. Beyond Frontal Faces: Improving Person Recognition Using Multiple Cues. In *CVPR*, 2015.
- [55] Yubo Zhang, Pavel Tokmakov, Cordelia Schmid, and Martial Hebert. A Structured Model For Action Detection. In *CVPR*, 2019.
- [56] Henry Zhou, Makarand Tapaswi, and Sanja Fidler. Now You Shake Me: Towards Automatic 4D Cinema. In *CVPR*, 2018.

We provide additional analysis of our task and models including confusion matrices, prediction examples for all our models, skewed distribution of number of samples for our classes, and diagrams depicting how we grouped the interaction and relationship classes.

A. Impact of Modalities

We analyze the impact of modalities by presenting qualitative examples where using multiple modalities help predict the correct interactions. Qualitative results presented here, refer to the quantitative performance indicated in Table 1. Fig. 7 shows that using dialog can help to improve predictions, Fig. 8 demonstrates the necessity of visual clip information and highlights that the two modalities are complementary. Finally, Fig. 9 shows that focusing on tracks (visual representations in which the two characters appear) provides further improvements to our model. Furthermore, Fig. 10 shows top-5 interaction classes that benefit most from using additional modalities.

Analyzing modalities. We also analyze the two models trained on only visual or only dialog cues (first two rows of Table 1). Some interactions can be recognized only with **visual** (v) features: *rides* 63% (v) / 0% (d), *walks* 29% (v) / 0% (d), *runs* 26% (v) / 0% (d); while others only with **dialog** (d) cues: *apologizes* 0% (v) / 66% (d), *compliments* 0% (v) / 26% (d), *agrees* 0% (v) / 25% (d).

Interactions that achieve non-zero accuracy with both modalities are: *hits* 64% (v) / 5% (d), *greet*s 12% (v) / 57% (d), *explains* 25% (v) / 51% (d).

Additionally, the top-5 predicted classes for **visual** cues are *asks* 77%, *hits* 64%, *rides* 63%, *watches* 49%, *talks on phone* 41%; and **dialog** cues are *asks* 75%, *apologizes* 66%, *greet*s 57%, *explains* 51%, *watches* 30%. As *asks* is the most common class, and *watches* is the second most common, these interactions work well with both modalities.

B. Joint Interaction and Relationships

Confusion matrices. Fig. 11 shows the confusion matrix in the top-15 most commonly occurring interactions on the validation and test sets. We see that multiple dialog based interactions (*e.g.* *talks to*, *informs*, and *explains*) are often confused. We also present confusion matrices for relationships in Fig. 12. A large part of the confusion is due to lack of sufficient data to model the tail of relationship classes.

Qualitative examples. Related to Table 2, Fig. 13 shows some examples where interaction predictions improve by jointly learning to model both interactions and relationships. Similarly, Fig. 14 shows how relationship classification benefits from our multi-task training setup.

C. Examples for *Who is Interacting*

Empirical evaluation shows that the knowledge about the relationship is important for localizing the pair of characters (Table 6). In Fig. 15, we illustrate an example where the dad walks into a room, sees his daughter *with* someone, and asks questions (see figure caption for details).

Finally, in Fig. 16, we show an example where the model is able to correctly predict all components (interaction class, relationship type and the pair of tracks) in a complex situation with more than 2 people appearing in the clip.

D. Dataset Analysis

Fig. 17 and Fig. 18 show normalized distributions for the number of samples in each class for train, validation and test sets of interactions and relationships respectively. As can be seen the most common classes appear many more times than the others. Data from a complete movie belongs to one of the three train/val/test sets to avoid model bias on the plot and characters behaviour. Notably, this means that the relative ratios between number of samples per class are also not necessarily consistent making the dataset and task even more challenging.

In the main paper, we described our approach to group over 300 interactions into 101 classes, and over 100 relationships into 15. We use radial tree diagrams to depict the groupings for interaction and relationship labels, visualized in Fig. 19 and 20 respectively.

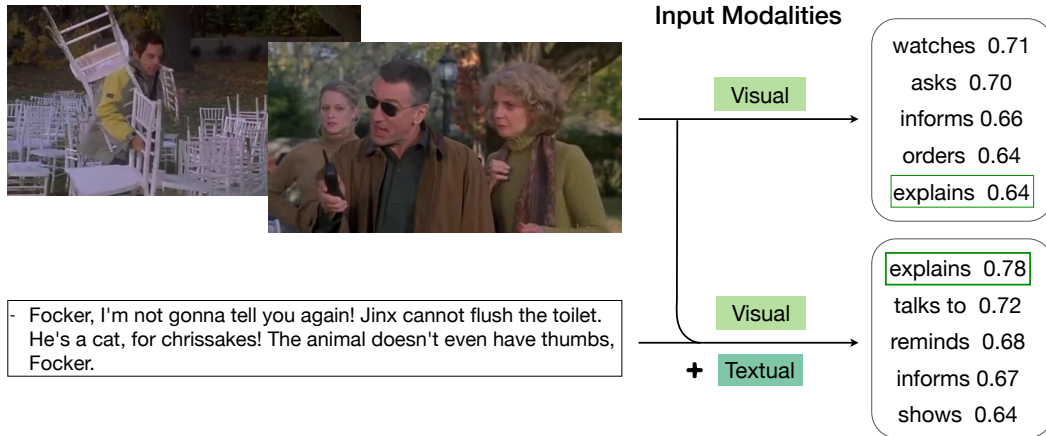


Figure 7: Improvement in prediction of **interactions** by including textual modality in addition to visual. The model learns to recognize subtle differences between interactions based on dialog. The example is from *Meet the Parents* (2000).

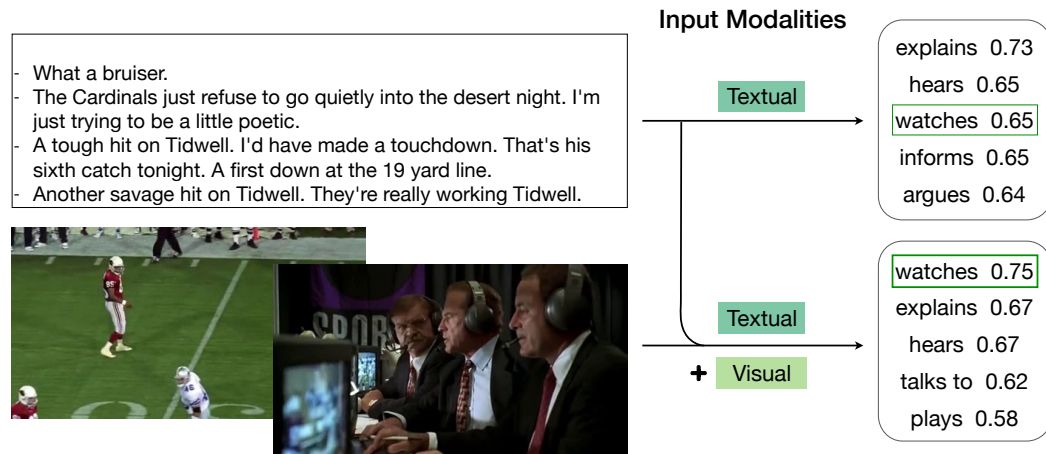


Figure 8: Improvement in prediction of **interactions** by including visual modality in addition to textual. The top-5 predicted interactions reflect the impact of visual input rather than relying only on the dialog. The example is from *Jerry Maguire* (1996).

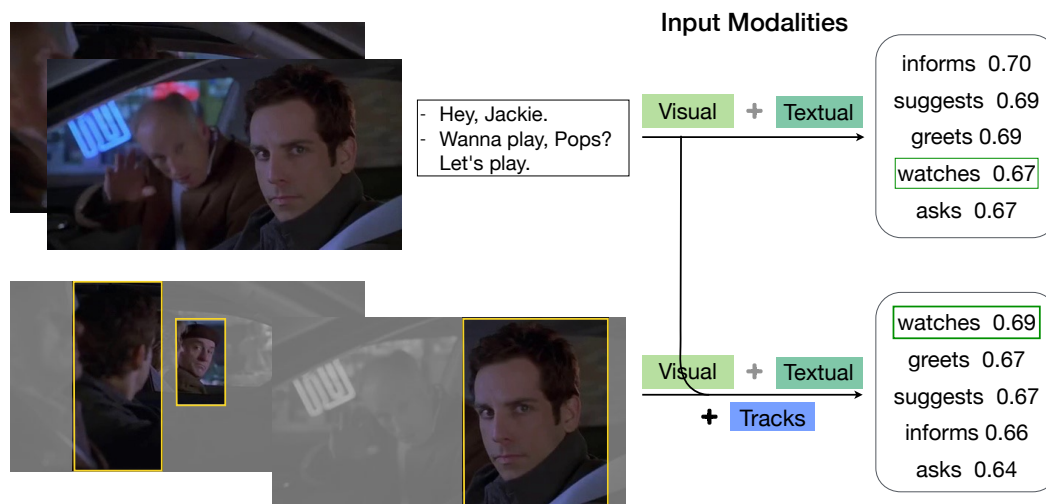


Figure 9: Improvement in prediction of **interactions** by including the pair of tracks modality in addition to visual and textual cues. The model can concentrate its attention on visual cues for the two people of interest instead of looking only at the clip level. The example is from *Meet the Parents* (2000).

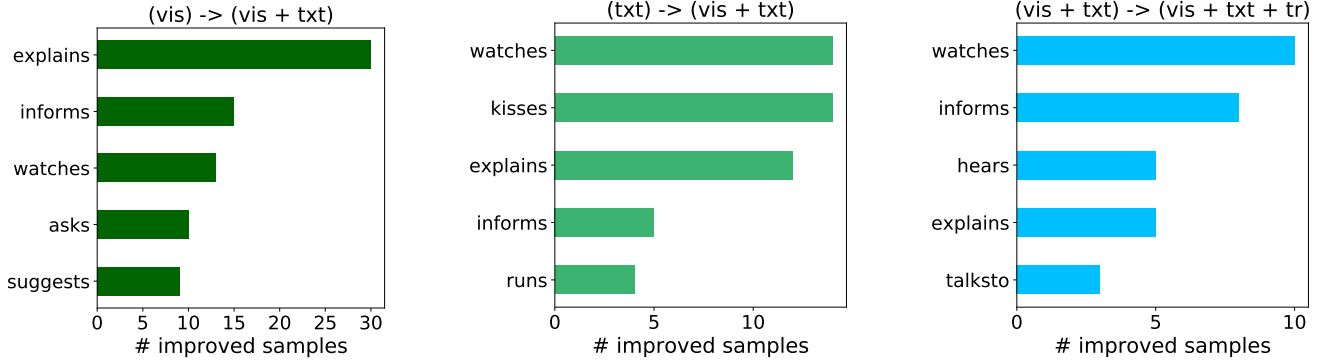


Figure 10: Each plot shows 5 **interaction** classes that have the most number of improved instances by including an additional modality. Specifically, the x -axis denotes the number of samples in which interaction prediction performance improves. **Left:** From only *visual* clip representation to *visual and textual*. As expected, using dialogues in addition to video frames boosts performance for classes that rely on dialog *e.g.* *explains*, *informs*. **Middle:** From only *textual* clip representation to *visual and textual*. Visual clip representations influence classes as *kisses*, *runs* during which people usually do not talk (dialog modality filled with zeros). **Right:** Finally, including all three modalities *visual, textual, tracks* improves performance over using *visual and textual*. Track pair localization improves recognition of classes typically used in group activities.

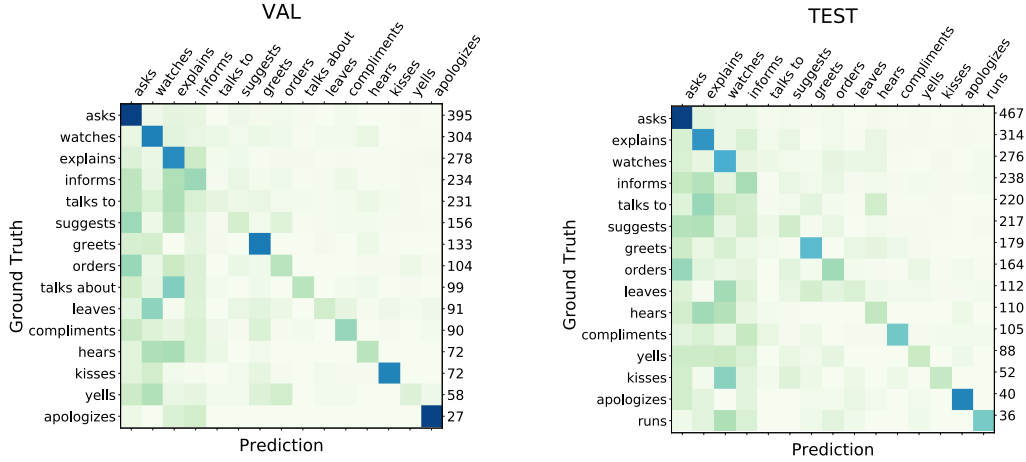


Figure 11: Confusion matrices for top-15 most common interactions for validation set (left) and test set (right). Model corresponds to the “Int. only” performance of 26.1% shown in Table 2. Numbers on the right axis indicate number of samples for each class

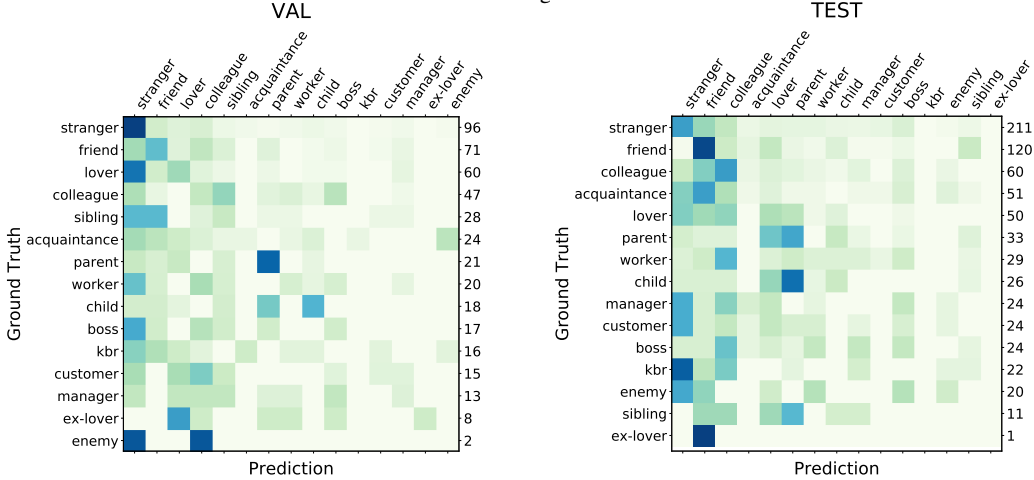


Figure 12: Confusion matrices for all relationships for validation set (left) and test set (right). Model corresponds to the “Rel. only” performance of 26.8% shown in Table 2. Numbers on the right axis indicate number of samples for each class.

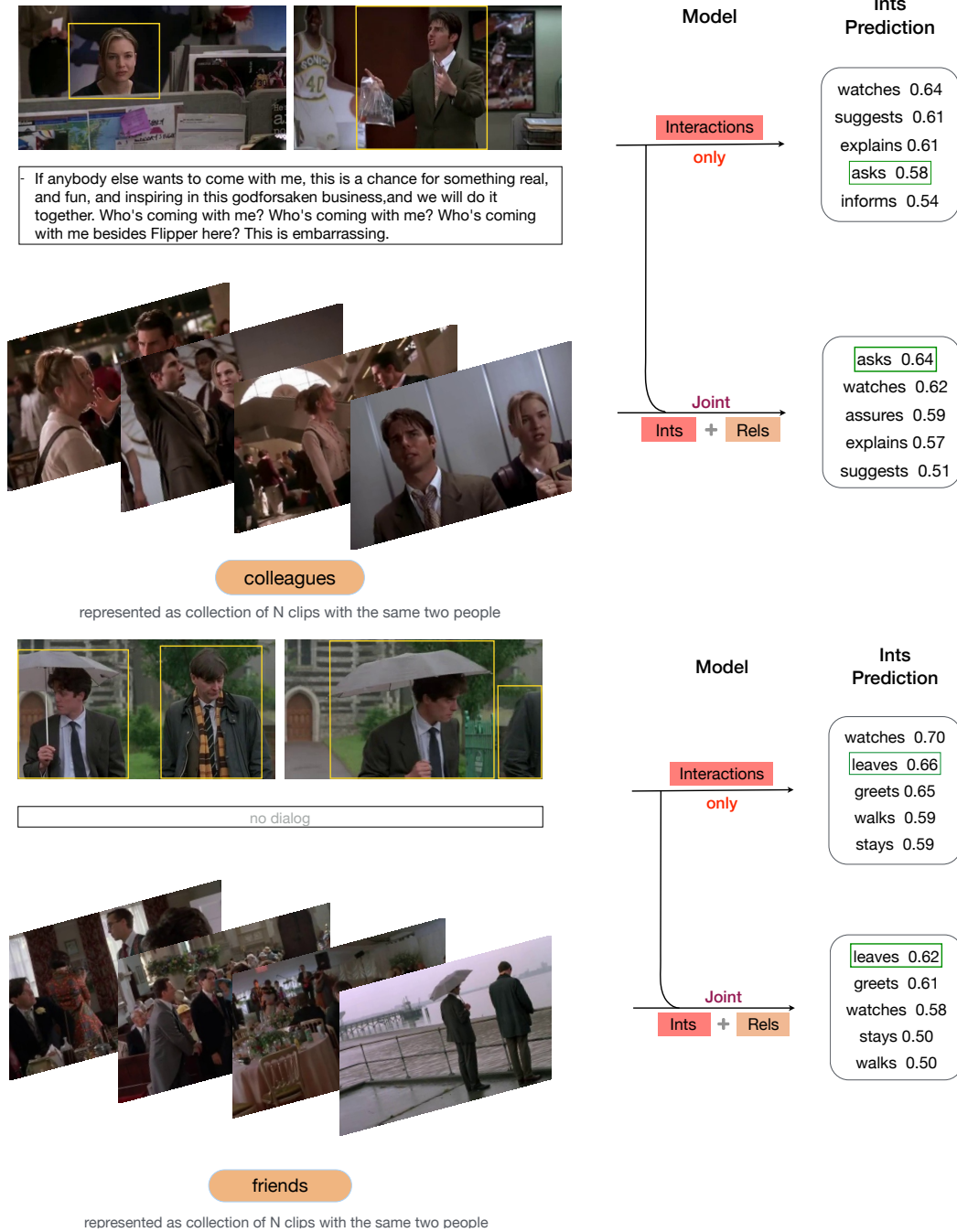


Figure 13: We show examples where training to predict interactions and relationships jointly helps improve the performance of **interactions**. **Top:** In the example from *Jerry Maguire* (1996), the joint model looks at several clips between Dorothy and Jerry and is able to reason about them being *colleagues*. This in turn helps refine the interaction prediction to *asks*. **Bottom:** In the example from *Four Weddings and Funeral* (1994), the model observes several clips from the entire movie where Charles and Tom are friends, and reasons that the interaction should be *leave* (which contains the *leave together* class). Note that there is no dialog for this clip.

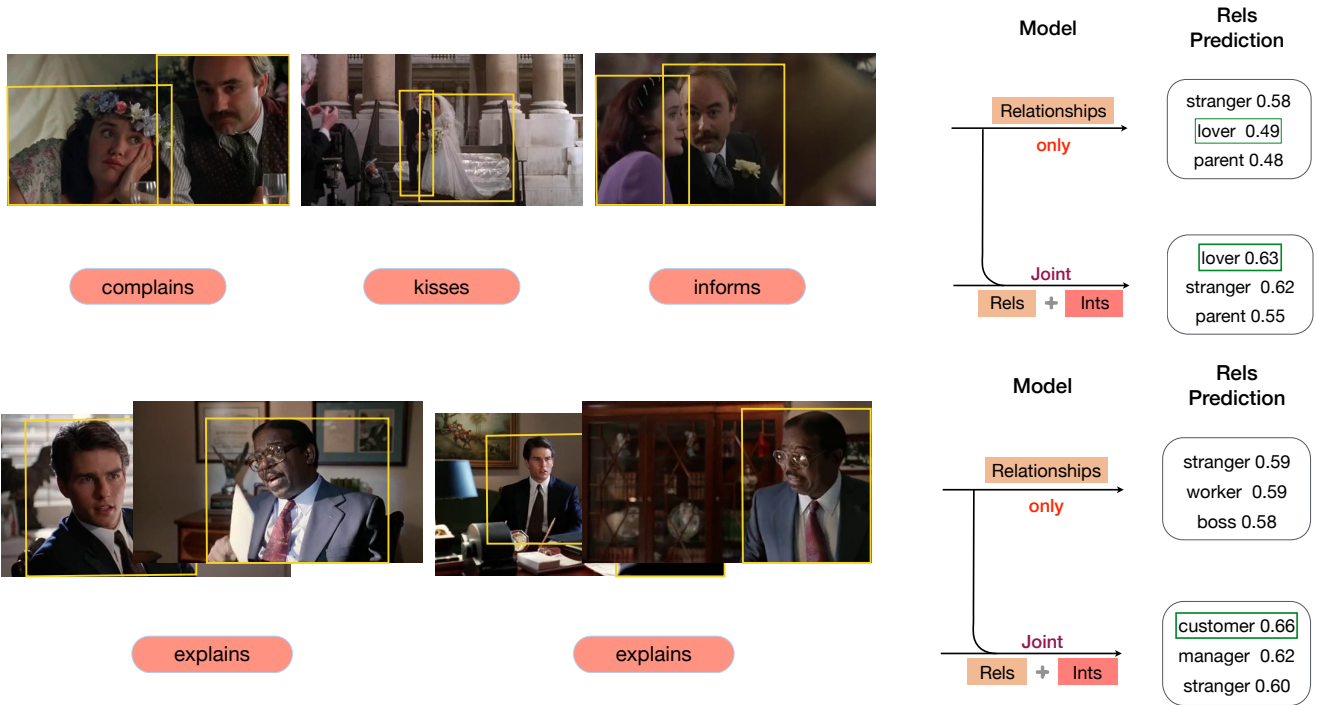


Figure 14: We show examples where training to predict interactions and relationships jointly helps improve the performance of **relationships**. **Top:** In the movie *Four Weddings and Funeral* (1994), clips between Bernard and Lydia exhibit a variety of interactions (e.g. kisses) that are more typical between *lovers* than *strangers*. **Bottom:** In the movie *The Firm* (1993), Frank and Mitch meet only once for a consultation, and are involved in two clips with the same interaction label *explains*. Our model is able to reason about this interaction, and it encourages the relationship to be *customer* and *manager*, instead of *stranger*.

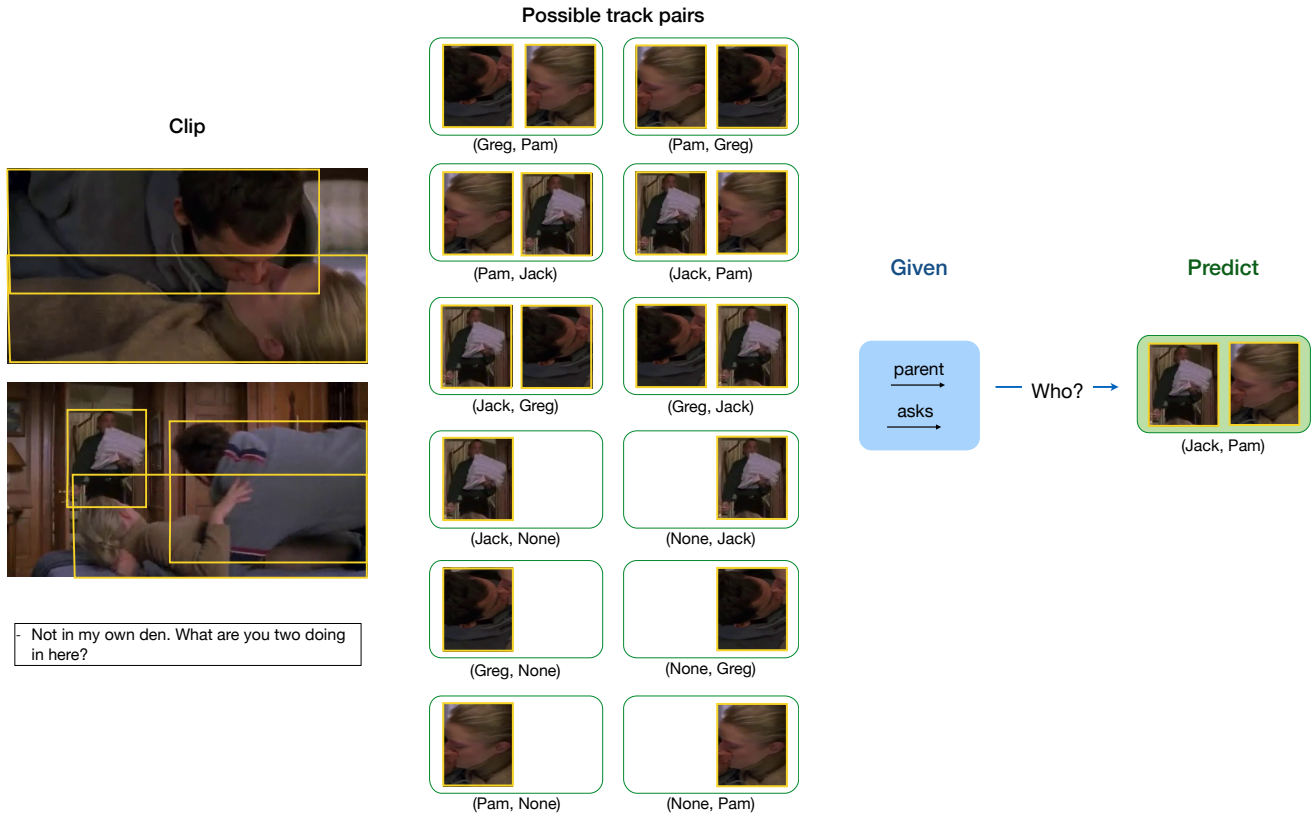


Figure 15: We illustrate an example from the movie *Meet the Parents* (2000) where a father (Jack) walks into a room while his daughter (Pam) and the guy (Greg) are kissing. Our goal is to predict the two characters when the interaction and relationship labels are provided. In this particular example, we see that Dad asks Pam a question (What are you two doing in here?). Note that their relationship is encoded as (Pam $\rightarrow$ child $\rightarrow$ Jack), or equivalently, (Jack $\rightarrow$ parent $\rightarrow$ Pam). When searching for the pair of characters with a given interaction **asks** and relationship as **parent**, our model is able to focus on the question at the clip level as it is asked by Jack in the interaction, and correctly predict *(Jack, Pam)* as the ordered character pair. Note that our model not only considers all possible directed track pairs (*e.g.* (Greg, Pam) and (Pam, Greg)) between characters, but also singleton tracks (*e.g.* (Jack, None)) to deal with situations when a person is absent due to failure in tracking or does not appear in the scene.

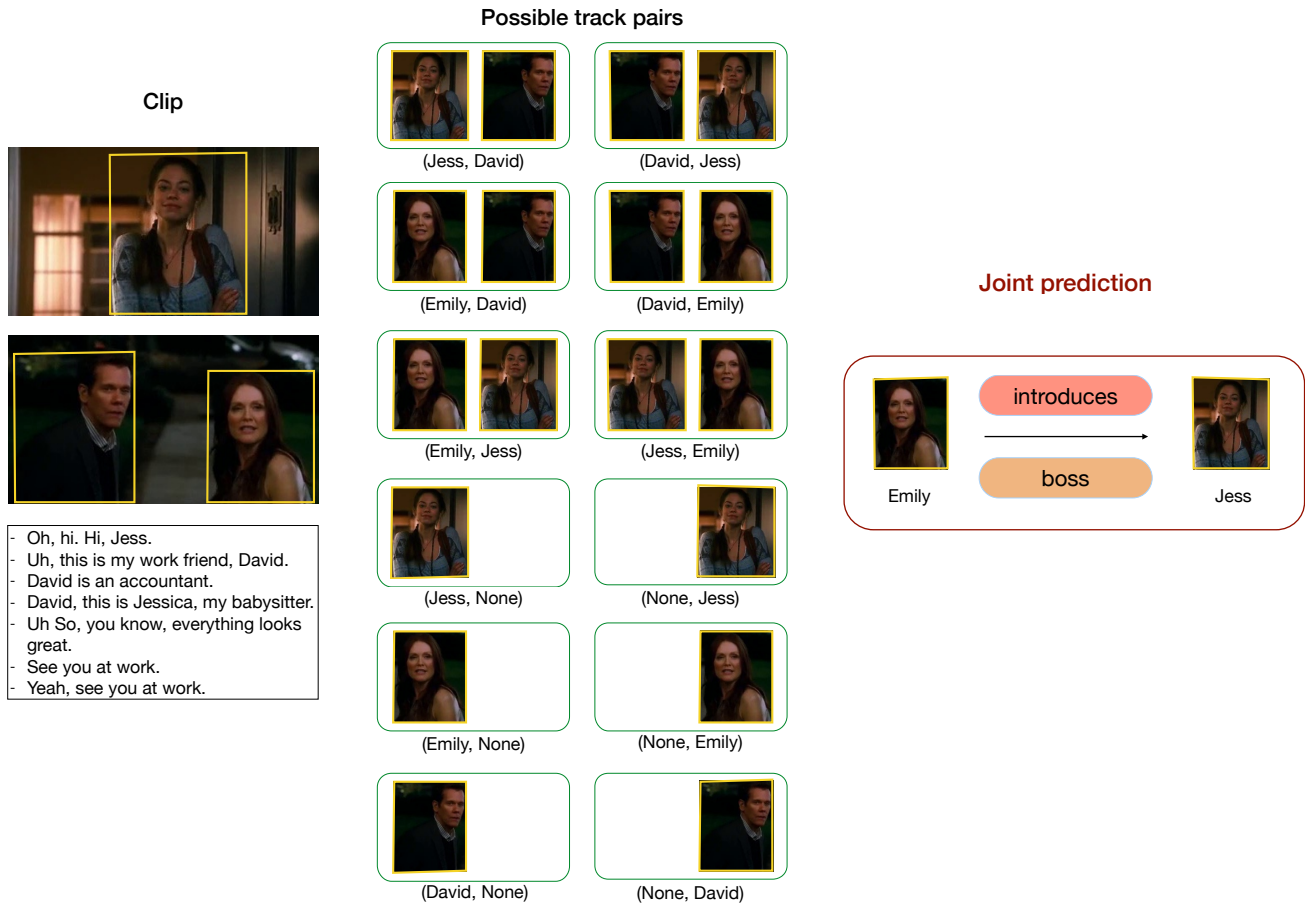


Figure 16: We present an example where our model is able to correctly and jointly predict all three components: track pair, interaction class and relationship type for the clip obtained from the movie *Crazy, Stupid, Love (2011)*. This clip contains three characters which leads to 12 possible track pairs (including singletons to deal with situations when a person is absent due to failure in tracking or does not appear in the scene). The model is able to correctly predict the two characters, their order, interaction and relationship. In this case, *Emily introduces David to Jess*. Jess is also her hired babysitter, and thus their relationship is – *Emily is boss of Jess*.

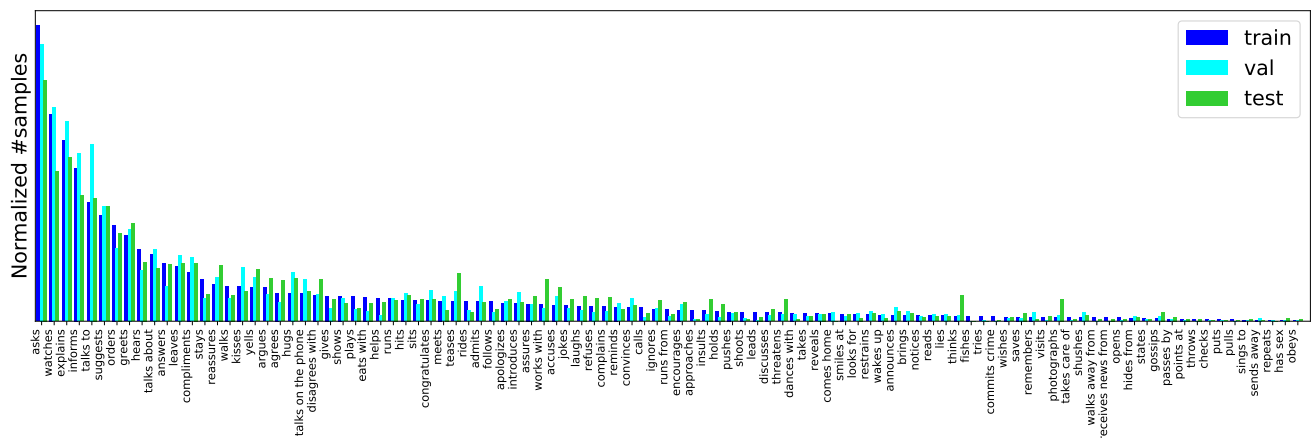


Figure 17: Distribution of interaction labels in train/val/test sets. Sorted by descending order based on train set.

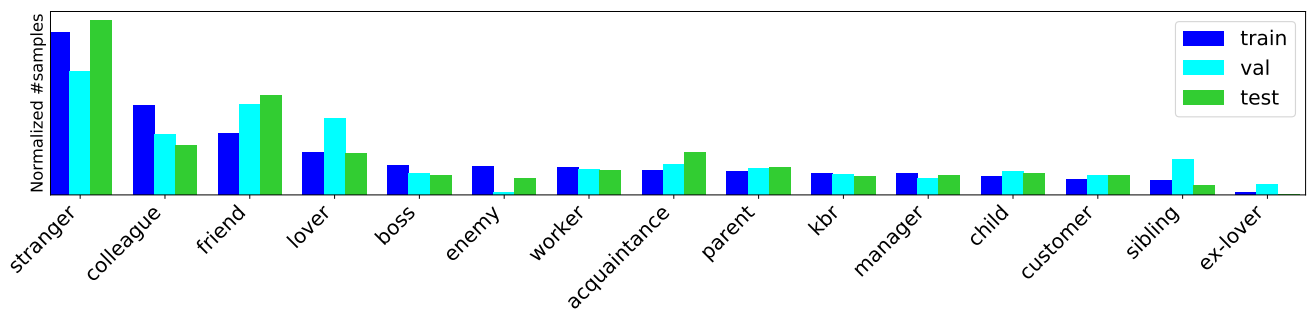


Figure 18: Distribution of relationship labels in train/val/test sets. Sorted by descending order based on train set.

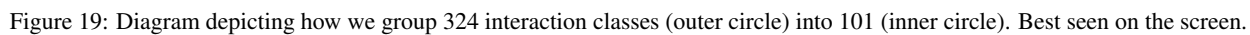


Figure 19: Diagram depicting how we group 324 interaction classes (outer circle) into 101 (inner circle). Best seen on the screen.

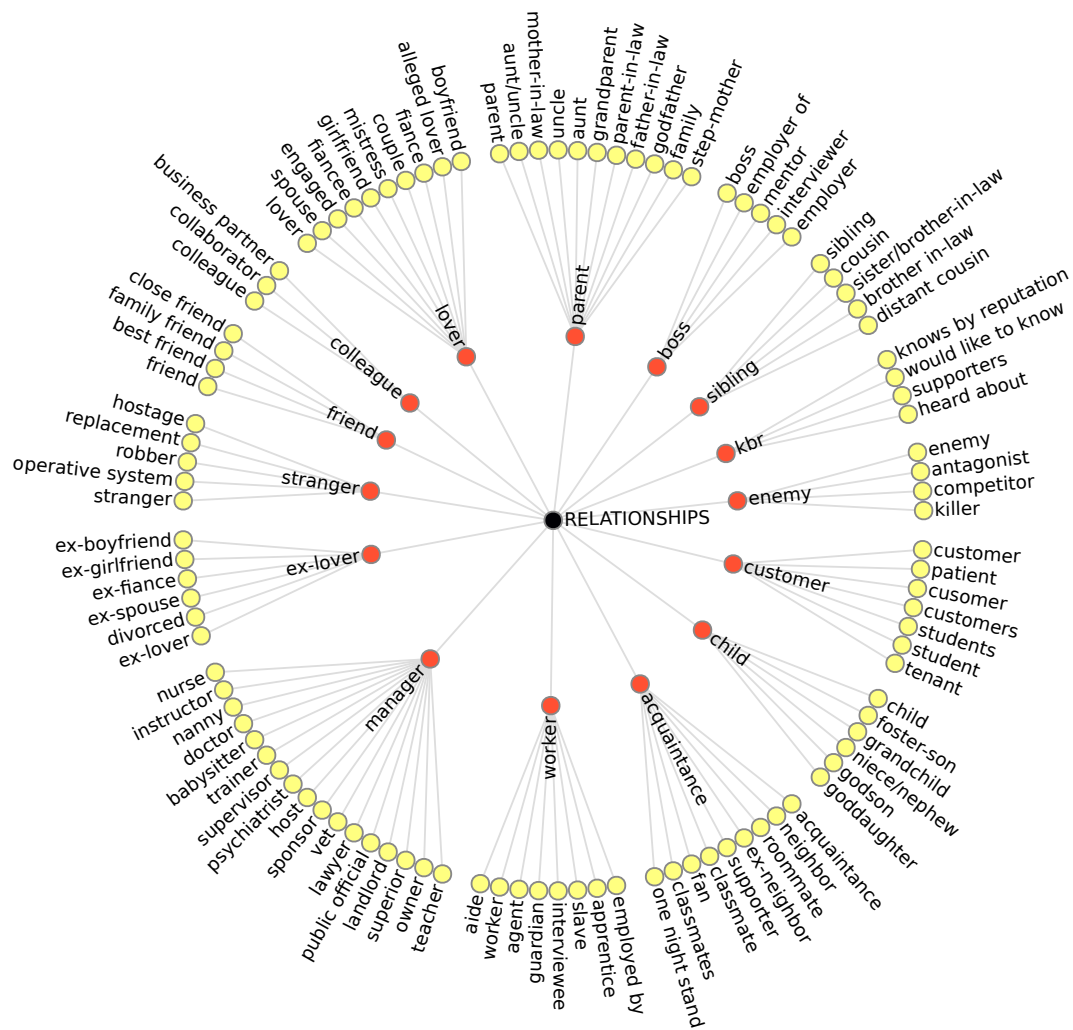


Figure 20: Diagram depicting how we group 107 relationship classes (outer circle) into 15 (inner circle).